From genes to protein structure and function: novel applications of computational approaches in the genomic era

Jeffrey Skolnick and Jacquelyn S. Fetrow

The genome-sequencing projects are providing a detailed 'parts list' of life. A key to comprehending this list is understanding the function of each gene and each protein at various levels. Sequence-based methods for function prediction are inadequate because of the multifunctional nature of proteins. However, just knowing the structure of the protein is also insufficient for prediction of multiple functional sites. Structural descriptors for protein functional sites are crucial for unlocking the secrets in both the sequence and structural-genomics projects.

Genome-sequencing projects are providing a detailed 'parts list' for life. Unfortunately, this list, a portion of which represents the amino acid sequence of all the proteins in a given genome, does not come with an instruction manual. That is, given the genome's sequences, one does not necessarily know straight away which regions encode proteins, which serve a regulatory role and which are responsible for the structure and replication of the DNA itself.

This is not unlike giving a child a list of parts necessary to create a working automobile. Without the necessary expertise, creating the final, working car from just the initial parts list is a nearly impossible task. Similarly, understanding how to create a complete, functioning cell given just the sequence of nucleotides found in an organism's genome is a complex problem.

What is a protein function?

After a genome is sequenced and its complete parts list determined, the next goal is to understand the function(s) of each part, including that of the proteins. What do we mean by protein function, the focus of this article?

Function has many meanings. At one level, the protein could be a globular protein, such as an enzyme, hormone or antibody, or it could be a structural or membrane-bound protein. Another level is its biochemical function, such as the chemical reaction and the substrate specificity of an enzyme. The regulatory molecules or cofactors that bind to a protein are also levels of biochemical function.

At the cellular level, the protein's function would involve its interaction with other macromolecules and the function and cellular location of such complexes. There is also the protein's physiological function; that is, in which metabolic pathway the protein is involved or what physiological role it performs in the organism. Finally, the phenotypic function is the role played by the protein in the total organism, which is observed by deleting or mutating the gene encoding the protein.

J. Skolnick (skolnick@danforthcenter.org) is at the Danforth Plant Science Center, Laboratory of Computational Genomics, 4041 Forest Park Avenue, St Louis, MO 63108, USA. J.S. Fetrow is at GeneFormatics, Suite 200, 5830 Oberlin Drive, San Diego, CA 92121-3754, USA.

Obviously, the complete characterization of protein function is difficult but efforts are under way at all levels[1–4], including cellular function[5,6]. In this article, however, we focus on identifying the biochemical function of a protein given its sequence, a problem that is amenable to molecular approaches.

Sequence-based approaches to function prediction

The sequence-to-function approach is the most commonly used function-prediction method. This robust field is well developed and, in the interest of space limitations, we will merely present a brief overview.

There are two main flavors of this approach: sequence alignment[7–9]; and sequence-motif methods such as Prosite[10], Blocks[11], Prints[12,13] and Emotif[14]. Both the alignment and the motif methods are powerful but a recent analysis has demonstrated their significant limitations[15], suggesting that these methods will increasingly fail as the protein-sequence databases become more diverse.

An extension of these approaches that combines protein-sequence with structural information has been developed and some successes have been reported[16]. However, this method still applies the structural information in a one-dimensional, 'sequence-like' fashion and fails to take into account the powerful three-dimensional information displayed by protein structures.

In addition, proteins can gain and lose function during evolution and may, indeed, have multiple functions in the cell (Box 1). Sequence-to-function methods cannot specifically identify these complexities. Inaccurate use of sequence-to-function methods has led to significant function-annotation errors in the sequence databases[17].

An alternative approach

An alternative, complementary approach to protein-function prediction uses the sequence-to-structure-to-function paradigm. Here, the goal is to determine the structure of the protein of interest and then to identify the functionally important residues in that structure. Using the chemical structure itself to identify functional sites is more in line with how the protein actually works.

0167-7799/00/$ – see front matter PII: S0167-7799(99)01398-0

In a sense, this is one long-term goal of 'structural genomics' projects[18,19], which are designed to determine all possible protein folds experimentally, just as genome-sequencing projects are determining all protein sequences[20]. This is in contrast to traditional structural-biology approaches, in which one knows the protein's function first and only then, if the function is sufficiently important, determines its structure.

It is implicitly assumed that having the protein's structure will provide insights into its function, thereby furthering the goals of the human-genome-sequencing project. However, knowing a protein's three-dimensional structure is insufficient to determine its function (Box 2). What we really need to analyse and predict the multifunctional aspects of proteins is a method specifically to recognize active sites and binding regions in these protein structures.

Active-site identification

In order to use a structure-based approach to function prediction, one must identify the key residues responsible for a given biochemical activity. For many years, it has been suggested that the active sites in proteins are better conserved than the overall fold. Taken to the limit, this suggests that one could not only identify distant ancestors with the same global fold and the same activity but also proteins with similar functions but distantly related, or possibly unrelated, global folds.

The validity of this suggestion was demonstrated empirically by Nussinov and co-workers, who showed that the active sites of eukaryotic serine proteases, subtilisins and sulfhydryl proteases exhibit similar structural motifs[21]. Furthermore, in a recent modeling study of *Saccharomyces cerevisiae* proteins, protein functional sites were found to be more conserved than other parts of the protein models[22]. Similarly, it has been demonstrated that the catalytic triad of the α/β hydrolases is structurally better conserved than other histidine-containing triads[23]. A comparison of the structure of the hydrolase catalytic triad to other histidine-containing triads shows a distinct bimodal distribution, while a similar analysis done with a randomly selected triad shows a unimodal distribution (Fig. 1).

Kasuya and Thornton[24] generalized this example by creating structural analogs of a few Prosite sequence motifs[10]. For the 20 most-frequently occurring Prosite patterns, the associated local structure is quite distinct. These results provide clear evidence that enzyme active sites are indeed more highly conserved than other parts of the protein.

Identifying active sites in experimental structures

Historically, several groups have attempted to identify functional sites in proteins; these efforts were directed at protein engineering or building functional sites in places where they did not previously exist. This has been successfully accomplished for several metal-binding sites[25–33]. However, highly accurate functional-site descriptors of the backbone and side-chain atoms were required, fueling the belief that significant atomic detail is required in site descriptors for function identification.

Highly detailed residue side-chain descriptors of the active sites of serine proteases and related proteins have been used to identify functional sites[3]. The use of these highly detailed motifs has led to the identification of several novel functional sites in known, high-quality protein structures[3,34]. More automated methods for finding spatial motifs in protein structures have also been described[21,34–40].

Unfortunately, most of these methods require the exact placement of atoms within protein backbones and side chains, and so have not been shown to be relevant to inexact predicted structures. Recently, however, we described the production of fuzzy, inexact descriptors of protein functional sites[15]. As we wish to apply the descriptors to experimental structures as well as to predicted protein models, we used only carbon atoms and side-chain centers-of-mass positions. We call these descriptors 'fuzzy functional forms' (FFFs) and have created them for both the disulfide-oxidoreductase[15,41] and α/β-hydrolase catalytic active sites[23].

The disulfide-oxidoreductase FFF was applied to screen high-resolution structures from the Brookhaven protein database[42]. In a dataset of 364 protein structures, the FFF accurately identified all proteins known to exhibit the disulfide-oxidoreductase active site[15]. In a larger dataset of 1501 proteins, the FFF again accurately identified all proteins with the active site. In addition, it identified another protein, 1fjm, a serine–threonine phosphatase. This result was initially discouraging but subsequent sequence alignment and clustering analysis strongly suggested that this putative site might indeed be a site of redox regulation in the serine–threonine phosphatase-1 subfamily[43]. If confirmed by experiment, this result will highlight the advantages of using structural descriptors to analyse multiple functional sites in proteins. It will also highlight the fact that human

Box 1. Proteins are multifunctional

A common protein characteristic that makes functional analysis based only on homology especially difficult is the tendency of proteins to be multifunctional. For instance, lactate dehydrogenase binds NAD, substrate and zinc, and performs a redox reaction. Each of these occurs at different functional sites that are in close proximity and the combination of all four sites creates the fully functional protein.

Other examples of multifunctional proteins are the nucleic-acid-binding proteins. For instance, DNA regulatory proteins often contain a DNA-binding domain, a multimerization domain and additional sites that bind regulatory proteins; a classic example is RecA[59]. The 3C rhinovirus protease exhibits a proteolytic function as well as an RNA-binding function[60,61]. Transcription factors are also complex, multifunctional proteins[62]. It is becoming increasingly important to recognize each of these different functions of gene products of a newly sequenced gene.

The serine–threonine-phosphatase superfamily is a prime example of the difficulties of using standard sequence analysis to recognize the multiple functions found in single proteins. This large protein family is divided into a number of subfamilies, all of which contain an essential phosphatase active site. Subfamilies 1, 2A and 2B exhibit 40% or more sequence identity between them[63]. However, each of these subfamilies is apparently regulated differently in the cell[64–67] and observation suggests that there are different functional sites at which regulation can occur. Because the sequence identity between subfamilies is so high, standard sequence-similarity methods could easily misclassify new sequences as members of the wrong subfamily if the functional sites are not carefully considered, as was recently demonstrated[43].

These are but a few examples of the multifunctionality of proteins. The recognition of this multifunctional nature is of critical importance to the genomics field. Useful functional-annotation methods must consider all of the specific functions in a given protein and will not just provide a general classification of function.

Box 2. Knowing a protein's structure does not necessarily tell you its function

Because proteins can have similar folds but different functions[68,69], determining the structure of a protein may or may not tell you something about its function. The most well-studied example is the $(\alpha/\beta)_8$ barrel enzymes, of which triose-phosphate isomerase (TIM) is the archetypal representative. Members of this family have similar overall structures but different functions, including different active sites, substrate specificities and cofactor requirements[70,71].

Is this example common? Our own analysis of the 1997 SCOP database[68] shows that the five largest fold families are the ferredoxin-like, the (α/β) barrels, the knottins, the immunoglobulin-like and the flavodoxin-like fold families with 22, 18, 13, 9 and 9 subfamilies, respectively (Fig. i). In fact, 57 of the SCOP fold families consist of multiple superfamilies. These data only show the tip of the iceberg, because each superfamily is further composed of protein families and each individual family can have radically different functions. For example, the ferredoxin-like superfamily contains families identified as Fe–S ferredoxins, ribosomal proteins, DNA-binding proteins and phosphatases, among others.

After this article was submitted, a much-more-detailed analysis of the SCOP database was published[72]. This finds a broad function–structure correlation for some structural classes, but also finds a number of ubiquitous functions and structures that occur across a number of families. The article provides a useful analysis of the confidence with which structure and function can be correlated[72]. Knowing the protein structure by itself is insufficient to annotate a number of functional classes and is also insufficient for annotating the specific details of protein function.

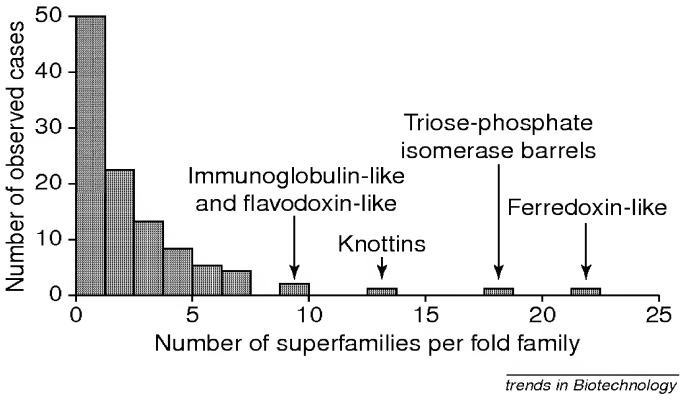

Figure i

Histogram of the numbers of superfamilies found in each SCOP fold family. These data clearly show that proteins with similar structures can have different functions and demonstrate the difficulty of assigning protein function based simply on the three-dimensional structure. The data were taken from the 1997 distribution of SCOP (http://scop.mrc-lmb.cam.ac.uk/scop). For a more-detailed analysis, see Ref. 72.

observation alone is no longer adequate for identifying all functional sites in known protein structures.

To date, the use of structure to identify function has largely focused on high-resolution structures and highly detailed descriptors of protein functional sites. However, the creation of inexact descriptors for functional sites opens the way to the application of these methods to inexact, predicted protein models. The question remains: how good does a model have to be in order to use FFFs to identify its active sites?

The state of the art in structure-prediction methods

For proteins whose sequence identity is above ~30%, one can use homology modeling to build the structure[44]. However, structure prediction is far more difficult for proteins that are not homologous to proteins with known structure. At present, there are two approaches for these sequences: *ab initio* folding[45–48] and threading[49–53].

In *ab initio* folding, one starts from a random conformation and then attempts to assemble the native structure. As this method does not rely on a library of pre-existing folds, it can be used to predict novel folds. The recent CASP3 protein-structure-prediction experiment (http://PredictionCenter.llnl.gov/CASP3) involved the blind prediction of the structure of proteins whose actual structure was about to be experimentally determined. These results indicate that considerable progress has been made[46,54]. For helical and α/β proteins with less than 110 residues, structures were often predicted whose backbone root-mean-square deviation (RMSD) from native ranged from 4–7 Å. Progress is being made with the β proteins, too, although they remain problematic. Because *ab initio* methods can identify novel folds, these methods could be used to help to select sequences likely to yield novel folds in experimental structural-genomics projects.

Another approach to tertiary-structure prediction is threading. Here, for the sequence of interest, one attempts to find the closest matching structure in a library of known folds[52,55]. Threading is applicable to proteins of up to 500 residues or so and is much faster than *ab initio* approaches. However, threading cannot be used to obtain novel folds.

Ab initio *predicted models can be used for automatic protein-function prediction*

The results of the recent CASP3 competition suggest that current modeling methods can often (but not always) create inexact protein models. Are these structures useful for identifying functional sites in proteins? Using the *ab initio* structure-prediction program MONSSTER, the tertiary structure of a glutaredoxin, 1ego, was predicted[56]. For the lowest-energy model, the overall backbone RMSD from the crystal structure was 5.7 Å.

To determine whether this inexact model could be used for function identification, the sets of correctly and incorrectly folded structures were screened with the FFF for disulfide-oxidoreductase activity[15]. The FFF uniquely identified the active site in the correctly folded structure but not in the incorrectly folded ones (Fig. 2). This is a proof-of-principle demonstration that inexact models produced by *ab initio* prediction of structure from sequence can be used for the subsequent prediction of biochemical function. Of course, improvements in the method have to be made before such predictions can be done on a routine basis.

Use of predicted structures from threading in protein-function prediction

At present, practical limitations preclude folding an entire genome of proteins using *ab initio* methods[57]. Threading is more appropriate for achieving the requisite high-throughput structure prediction. Thus, a standard threading algorithm[58] has been used to screen all

proteins in nine genomes for the disulfide-oxidoreductase active site described above.

First, sequences that aligned with the structures of known disulfide oxidoreductases were identified. Then, the structure was searched for matches to the active-site residues and geometry. For those sequences for which other homologs were available, a sequence-conservation profile was constructed[23]. If the putative active-site residues were not conserved in the sequence subfamily to which the protein belongs, that sequence was eliminated. Otherwise, the sequence is predicted to have the function.

Using this sequence-to-structure-to-function method, 99% of the proteins in the nine genomes that have known disulfide-oxidoreductase activity have been found. From 10% to 30% more functional predictions are made than by alternative sequence-based approaches; similar results are seen for the α/β hydrolases[23]. Surprisingly, in spite of the fact that threading algorithms have problems generating good sequence-to-structure alignments, active sites are often accurately aligned, even for very distant matches. This observation would agree with the above experimental results indicating that active sites are well conserved in protein structures.

Importantly, the false-positive rate when using structural information is much lower than that found using sequence-based approaches, as demonstrated by a detailed comparison of the FFF structural approach and the Blocks sequence-motif approach (N. Siew *et al.*, unpublished). In this study, the sequences in eight genomes, including *Bacillus subtilis*, were analysed for disulfide-oxidoreductase function using the disulfide-oxidoreductase FFF, the thioredoxin Block 00194 and the glutaredoxin Block 00195. If we assume that those sequences identified by both the FFF and Blocks are 'true positives', we find 13 such sequences in the *B. subtilis* genome.

There is no experimental evidence validating all of these 'true positives' and so they are more accurately termed 'consensus positives'. In order to find these 13 'consensus positive' sequences, the FFF hits seven false positives. On the other hand, Blocks hits 23 false positives (Fig. 3). It was previously suggested that the use of a functional requirement adds information to threading and reduces the number of false positives[52]. These data, including the data shown in Fig. 3, validate this claim on a genome-wide basis.

Of course, as no genome has had the function of all of its proteins experimentally annotated, it is impossible to know how many other proteins with the specified biochemical function were not properly identified. This is a critical question for researchers attempting to predict protein function. Experimental confirmation will be needed to validate this or any other method fully. This points out the need for closely coupling computational function-prediction algorithms with experiments.

Weaknesses of using the sequence-to-structure-to-function method of function prediction

Based on studies to date, the identification of enzymatic activity requires a model in which the backbone RMSD from native near the active sites is about 4–5 Å. Predicted models are better at describing the geometry in the core of the molecule than in the loops and so predicting the function of a protein whose active site is in loops may be a problem. Also, the method can currently only be applied to enzyme active sites; substrate- and ligand-binding sites have not been identified using the inexact models. Techniques that will further refine inexact protein models will be quite useful in taking the protein analysis to the next step.

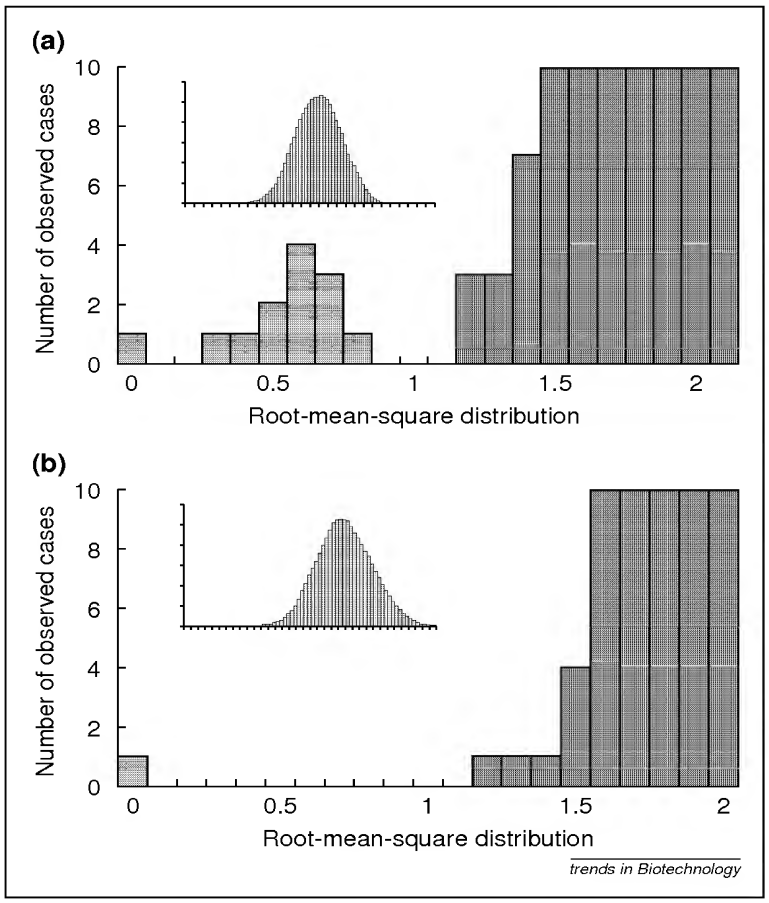

Figure 1

The distribution of root-mean-square distributions (RMSD) between the hydrolase catalytic triad and all other histidine-containing triads shows a bimodal distribution **(a)**; by contrast, the RMSD between a randomly selected (non-catalytic) triad and all other histidine-containing triads has a unimodal distribution **(b)**. The His–Ser–Asp catalytic trial in the protein-1 gpI (Rp2 lipase) **(a)** and a random histidine-containing triad from 4pga (glutaminase–asparaginase) **(b)** were structurally aligned to all His-containing triads in a database of 1037 proteins[23]. Actual α/β-hydrolase active sites **(a)** and the 4pga site **(b)** are indicated by blue bars; other histidine triads that are not active sites are indicated by red bars. None of the sites found by matching to the 4pga were hydrolase active sites. Inset graphs show the full distribution.

Conclusions

Although sequence-based approaches to protein-function prediction have proved to be very useful, alternatives are needed to assign the biochemical function of the 30–50% of proteins whose function cannot be assigned by any current methods. One emerging approach involves the sequence-to-structure-to-function paradigm. Such structures might be provided by structural-genomics projects or by structure-prediction algorithms. Functional assignment is made by screening the resulting structure against a library of structural descriptors for known active sites or binding regions.

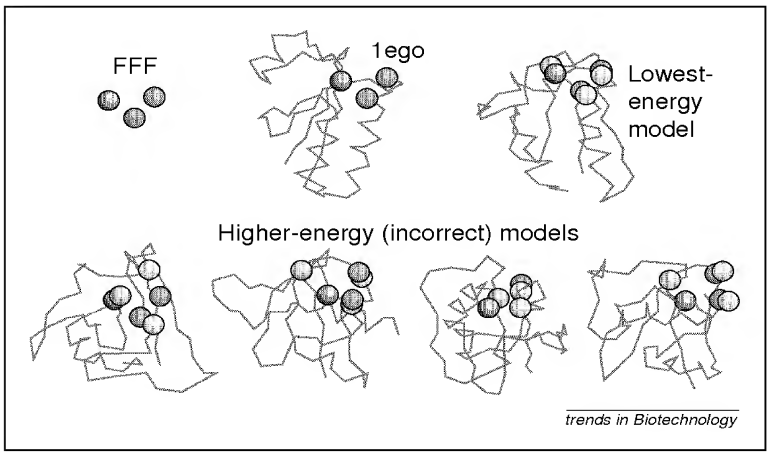

Figure 2

Application of the disulfide-oxidoreductase fuzzy functional form (FFF) to *ab initio* models of glutaredoxin created by the program MONSSTER shows that the FFF can distinguish between correctly folded and misfolded (or higher-energy) models. The FFF is shown as two orange balls (representing the cysteines) and a blue ball (representing the proline). The protein models are shown as magenta wire models with the active-site cysteines and proline shown as yellow and cyan balls, respectively. The FFF clearly distinguishes the correct active site in the crystal structure of the glutaredoxin 1ego and the correctly folded, lowest-energy model. The FFF does not match to the active sites of any of the higher energy, misfolded structures, four of which are shown here.

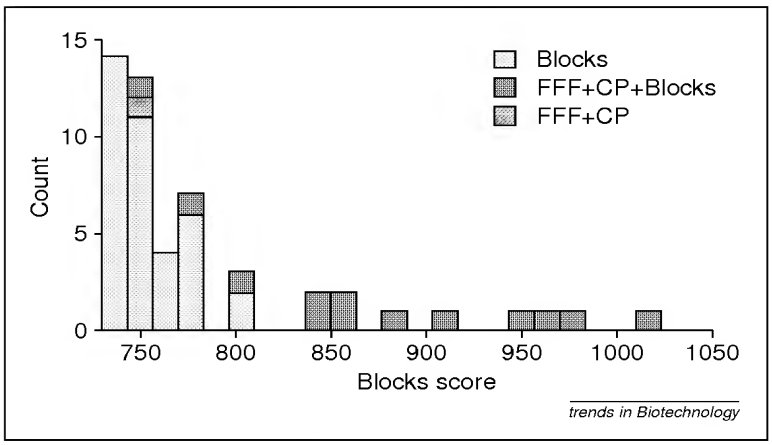

Figure 3

Analysis of the *Bacillus subtilis* genome using the thioredoxin Block 00194. The Blocks score (computed using the publicly available BLIMPS program) is plotted on the *x* axis and the number of sequences found in each scoring bin is plotted on the *y* axis. Those sequences identified as 'consensus positives' [identified by both the fuzzy functional form (FFF) and the Block] are shown as red bars. One additional sequence found by the FFF, which is likely to be a true positive, is shown as a blue bar. All other sequences, putative 'false positives', are shown as yellow bars. Using the Blocks score at which all 13 of the 'consensus positives' are found, 23 false positives are also found. In its analysis of the *B. subtilis* genome, the FFF identifies only seven false positives along with the same 13 'consensus positives' (data not shown).

Detailed descriptors will only work on the experimentally determined, high-quality structures. Ideally, however, the descriptors should work on both experimental structures and the cruder models provided by tertiary-structure-prediction algorithms.

The advantages of such an approach are that one need not establish an evolutionary relationship in order to assign function, that more than one function can be assigned to a given protein [an issue of major importance, because proteins are multifunctional (Box 1)] and, ultimately, that having a structure can provide deeper insight into the biological mechanism of protein function and regulation. The disadvantages are that one needs to have the protein's structure before a function can be assigned and that the approach is limited to those functions associated with proteins with at least one solved structure, so that a functional-site descriptor can be constructed.

In this sense, structure-to-function assignment can be thought of as 'functional threading' – find the active-site match in a library of descriptors for known protein active sites. This is the first step in the long process of using structure to assign all levels of function, a goal that is made increasingly important with the emergence of structural genomics. Based on the progress to date, it is apparent that structure will play an important role in the post-genomic era of biology.

Acknowledgment

We thank L. Zhang for producing the data in Box 2 and Fig. 1.

References

1 Gurd, F.R.N. and Rothgeb, T.M. (1979) Motions in proteins. *Adv. Protein Chem.* 33, 73–165
2 Laskowski, R.A. *et al.* (1996) X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.* 259, 175–201
3 Wallace, A.C. *et al.* (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser–His–Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 5, 1001–1013
4 Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19, 6565–6572
5 Riley, M. (1993) Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952
6 Karp, P.D. and Riley, M. (1993) Representations of metabolic knowledge. *Ismb* 1, 207–215
7 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
8 Pearson, W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.* 266, 227–258
9 Sturrock, S.S. and Collins, J.F. (1993) *Biocomputing Research Unit*, University of Edinburgh, Edinburgh, UK
10 Bairoch, A. *et al.* (1995) The PROSITE database, its status in 1995. *Nucleic Acids Res.* 24, 189–196
11 Henikoff, S. and Henikoff, J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics* 19, 97–107
12 Attwood, T.K. *et al.* (1994) PRINTS – A database of protein motif fingerprints. *Nucleic Acids Res.* 22, 3590–3596
13 Attwood, T.K. *et al.* (1997) Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res.* 25, 212–216
14 Nevill-Manning, C.G. *et al.* (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5865–5871
15 Fetrow, J.S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949–968
16 Yu, L. *et al.* (1998) A homology identification method that combines protein sequence and structure information. *Protein Sci.* 7, 2499–2510
17 Bork, P. and Bairoch, A. (1996) Go hunting in sequence databases but watch out for traps. *Trends Genet.* 12, 425–427
18 Gaasterland, T. (1998) Structural genomics: bioinformatics in the driver's seat. *Nat. Biotechnol.* 16, 625–627
19 McKusick, V.A. (1997) Genomics: structural and functional studies of genomes. *Genomics* 45, 244–249
20 Montelione, G.T. and Anderson, S. (1999) Structural genomics: keystone for a human proteome project. *Nat. Struct. Biol.* 6, 11–12

21 Fischer, D. *et al.* (1994) Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* 3, 769–778

22 Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13597–13602

23 Zhang, L. *et al.* (1998) Functional analysis of *E. coli* proteins for members of the α/β hydrolase family. *Fold. Design* 3, 535–548

24 Kasuya, A. and Thornton, J.M. (1999) Three-dimensional structure analysis of Prosite patterns. *J. Mol. Biol.* 286, 1673–1691

25 Coldren, C.D. *et al.* (1997) The rational design and construction of a cuboidal iron–sulfur protein. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6635–6640

26 Pinto, A.L. *et al.* (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5562–5567

27 Hellinga, H.W. and Richards, F.M. (1991) Construction of new ligand binding sites in proteins of known structure: (I) computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* 222, 763–785

28 Hellinga, H.W. *et al.* (1991) Construction of new ligand binding sites in proteins of known structure: (II) grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* 222, 787–803

29 Klemba, M. and Regan, L. (1995) Characterization of metal binding by a designed protein: single ligand substitutions at a tetrahedral Cys_2His_2 site. *Biochemistry* 34, 10094–10100

30 Klemba, M. *et al.* (1995) Novel metal-binding proteins by design. *Nat. Struct. Biol.* 2, 368–373

31 Farinas, E. and Regan, L. (1998) The *de novo* design of a rubredoxin-like Fe site. *Protein Sci.* 7, 1939–1946

32 Crowder, M.W. *et al.* (1995) Spectroscopic studies on the designed metal-binding sites of the 43C9 single chain antibody. *J. Am. Chem. Soc.* 117, 5627–5634

33 Halfon, S. and Craik, C.S. (1996) Regulation of proteolytic activity by engineered tridentate metal binding loops. *J. Am. Chem. Soc.* 118, 1227–1228

34 Wallace, A.C. *et al.* (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. *Protein Sci.* 6, 2308–2323

35 Kleywegt, G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285, 1887–1897

36 Matsuo, Y. and Nishikawa, K. (1994) Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci.* 3, 2055–2063

37 Russell, R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* 279, 1211–1227

38 Han, K.F. *et al.* (1997) Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* 6, 1587–1590

39 Artymiuk, P.J. *et al.* (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* 236, 327–344

40 Karlin, S. and Zhu, Z.Y. (1996) Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.* 93, 8344–8349

41 Fetrow, J.S. *et al.* (1998) Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703–711

42 Abola, E.E. *et al.* (1987) *Protein Data Bank in Crystallographic Databases: Information Content, Software Systems, Scientific Application* (Allen, F.H. *et al.*, eds), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester

43 Fetrow, J.S. *et al.* (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* 13, 1866–1874

44 Sali, A. *et al.* (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins* 23, 318–326

45 Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281, 565–577

46 Shortle, D. (1999) The state of the art. *Curr. Biol.* 9, R205–R209

47 Lee, J. *et al.* (1999) Calculation of protein conformation by global optimization of a potential energy function. *Proteins* 3 (Suppl.), 204–208

48 Ortiz, A. *et al.* (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins* 3 (Suppl.), 177–185

49 Bowie, J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170

50 Finkelstein, A.V. and Reva, B.A. (1991) A search for the most stable folds of protein chains. *Nature* 351, 497–499

51 Bryant, S.H. and Lawrence, C.E. (1993) An empirical energy function for threading protein sequence through folding motif. *Proteins* 16, 92–112

52 Lathrop, R. and Smith, T.F. (1996) Global optimum protein threading with gapped alignment and empirical pair scoring function. *J. Mol. Biol.* 255, 641–665

53 Jones, D.T. *et al* (1992). A new approach to protein fold recognition. *Nature* 358, 86–89

54 Sternberg, M.J. *et al.* (1999) Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* 9, 368–373

55 Miller, R.T. *et al.* (1996) Protein fold recognition by sequence threading tools and assessment techniques. *FASEB J.* 10, 171–178

56 Ortiz, A.R. *et al.* (1998) Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* 277, 419–448

57 Skolnick, J. *et al.* (1998) Reduced protein models and their application to the protein folding problem. *J. Biomol. Struct. Dyn.* 16, 381–396

58 Jaroszewski, L. *et al.* (1998) Fold prediction by a hierarchy of sequence, threading and modeling methods. *Protein Sci.* 7, 1431–1440

59 Takahashi, M. *et al.* (1996) Locations of functional domains in the RecA protein: overlap of domains and regulation of activities. *Eur. J. Biochem.* 242, 20–28

60 Leong, L.E. *et al.* (1993) Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5′ noncoding region of the viral RNA: evidence that 3Cpro has different domains for the RNA binding and proteolytic activities. *J. Biol. Chem.* 268, 25735–25739

61 Matthews, D.A. *et al.* (1994) Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site and means for cleaving precursor polyprotein. *Cell* 77, 761–771

62 Ladomery, M. (1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *BioEssays* 19, 903–909

63 Goldberg, J. *et al.* (1995) Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* 376, 745–753

64 Mumby, M.C. and Walter, G. (1993) Protein serine/threonine phosphatases: structure, regulation and functions in cell growth. *Physiol. Rev.* 73, 673–699

65 Jia, Z. (1997) Protein phosphatases: structures and implications. *Biochem. Cell Biol.* 75, 17–26

66 Holmes, C.F.B. and Boland, M.P. (1993) Inhibitors of protein phosphatase-1 and -2A: two of the major serine/threonine protein phosphatases involved in cellular regulation. *Curr. Opin. Struct. Biol.* 3, 934–943

67 Nemani, R. and Lee, E.Y.C. (1993) Reactivity of sulfhydryl groups of the catalytic subunits of rabbit skeletal muscle protein phosphatases 1 and 2A. *Arch. Biochem. Biophys.* 300, 24–29

68 Murzin, A.G. *et al.* (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

69 Orengo, C.A. *et al.* (1997) CATH: a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108

70 Lesk, A.M. *et al.* (1989) Structural principles of α/β proteins: the packing of the interior of the sheet. *Proteins Struct. Funct. Genet.* 5, 139–148

71 Farber, G.K. and Petsko, G.A. (1990) The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* 15, 228–234

72 Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288, 147–164